Deep Structured Network with Joint and Interpretable Bottom-up and Top-down Inference

Wenguan Wang*[1,2], Yuanlu Xu*[2], Quanshi Zhang[2], Jianbing Shen†[1], and Song-Chun Zhu[2]

[2]UCLA, Technical Report #190003

[1]Beijing Lab of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, China
[2]Department of Computer Science and Statistics, University of California, Los Angeles, USA

Abstract

This paper proposes a structured neural network that combines the learning power of deep learning and the interpretable structured representation of graphical models. The proposed deep hierarchical model, i.e., α-β-γ network, not only explores how hierarchical graphical structures are represented in neural network, but also focuses on how predictions are conducted. In particular, with the direct application of modern network architectures, three kinds of information flows, from image input to label output (i.e., straight pass), low level to high level (i.e., bottom-up process), high level to low level (i.e., top-down process), are integrated and learned in end-to-end and back-propagation manner. We apply the proposed model in two example tasks, e.g., cloth landmark estimation and human pose detection and demonstrate the performance improvements with joint bottom-up and top-down inference. More importantly, the proposed network confers interpretability on how each flow contributes to the final prediction, where post-hoc explanations coincide with human intelligence.

1. Introduction

Neural networks are currently revolutionizing computer vision. Their wide-ranging success has proven their strong representation power and end-to-end learning ability. However, they may not directly encode interpretable structures and top-down information. For example, it's difficult to incorporate the knowledge of human body decomposability into networks since the intrinsic mechanism of a network is often hard to explain. Alternatively, graphical models are powerful to build structured representations, which is the incentive for their prevalence in computer vision. Such structured representations could reflect task-specific relations and constraints. For example, in cloth landmark localization (see Fig. 1), nodes represent atomic components (*e.g.*, collars, hems, *etc.*), and edges describe node interrelations (*e.g.*, kinematic dependencies among cloth landmarks). Graphical models allow domain experts to inject their high-level knowledge, but often require significant feature engineering.

We propose a deep structured network, named α-β-γ network, which augments the hierarchical graphical representation with the learning capability of neural network, and pursue to connote three information flows, straight pass (*i.e.*, α process), bottom-up process (*i.e.*, β process) and top-down process (*i.e.*, γ process), in hierarchical models. As illustrated in Fig. 1, when predicting the location of *upper-body cloth* of a person, we consider three kinds of information: image regions directly revealing itself, the decompositional relation from parent *full-body cloth*, and the compositional relations from children *l.&r. collar* and *l.&r. hem*. Different information flows confer different portions of contributions to the final prediction. There are some interesting properties about α-β-γ network:

• **Encodings of joint bottom-up and top-down inference**. Our structured network models graph nodes as CNNs and takes into account the dependencies (*e.g.*, composition, decomposition, and contextual relation) within the hierarchical graph. It provides a principled algorithm for learning hierarchical graphical models jointly with neural networks. The proposed model approaches three basic inference processes [34] with end-to-end learning: α process directly generates predictions based on image features, β process

*Wenguan Wang and Yuanlu Xu contributed equally.

†Corresponding author: Jianbing Shen (shenjianbing@bit.edu.cn).
This work was supported by the Beijing Natural Science Foundation under Grant 4182056, the Fok Ying Tung Education Foundation under Grant 141067, and the Specialized Fund for Joint Building Program of Beijing Municipal Education Commission. This work was also supported by ONR MURI project N00014-16-1-2007, DARPA XAI Award N66001-17-2-4029, and NSF IIS 1423305.

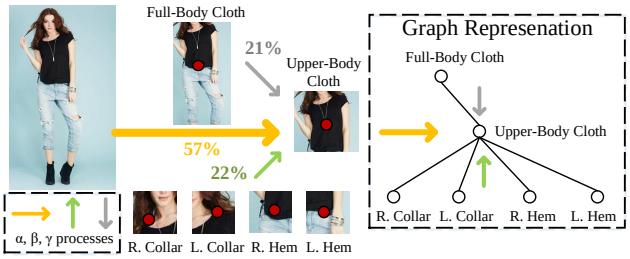


Figure 1. **Illustration of joint bottom-up and top-down inference**. Given a hierarchy of human cloth, three types of information (*i.e.*, α, β, and γ processes) contribute to the final prediction of *upper body cloth* node. With α-β-γ network, these three processes can be explicitly learnt in an end-to-end manner with post-hoc interpretability.

makes predictions by binding child node(s) in bottom-up fashion and γ process utilizes contextual information from parent node(s) in top-down style. We show that, with the suggested model, the hierarchical graph with above three processes can be efficiently learned in a stochastic way.

• **Post-hoc interpretability**. A major benefit of our network lies in the post-hoc interpretability. Taking human pose detection as an example, it is intuitive that people usually directly observe a certain node (*e.g.*, human arm), without occlusion. When the arm node is partially occluded, people rely more on bottom-up process that considers the information from those non-occluded child nodes (*e.g.*, hand). When the arm node is heavily occluded or becomes indistinguishable, people would still recognize this node with the high-level prior knowledge of human body articulation. As seen, α, β and γ processes are straightforward and interpretable, which leads to more interpretability compared with previous structured deep learning methods. Thus our model is able to provide information that can be interpreted by humans as how results are inferred and combined from top-down/bottom processes. Additionally, such self-interpretability is also measurable, which can be evaluated as the agreement between inference processes and respective human performance.

We conduct experiments on two tasks, *i.e.*, cloth landmark localization and human pose detection, to verify the effectiveness and generalization of our model. The selected two tasks by nature implies complex hierarchical structures. Results show that our method outperforms competing methods *with the similar model complexity*. From the experimental results, we further observe that: (i) α process is generally stronger than β and γ processes; (ii) α process is favored for low-level nodes (*e.g.*, cloth landmarks, human joints), while β and γ processes are preferred for high-level nodes (*e.g.*, full-body cloth, or upper-body pose); (iii) combining three inference processes is beneficial to final predictions.

Contributions. The contributions are three-field: i) a deep network representing hierarchical graphical structures; ii) explicit encodings of three inference processes with end-to-end learning; iii) post-hoc interpretability.

The rest of this paper is organized as follows. We first review the related work in § 2, then discuss the representation and formulation of our model in § 3 and § 4, respectively. We further elaborate the learning process in § 5. We report experiments and comparisons in § 6, and finally conclude this paper in § 7.

2. Related Work

We give a categorized overview of the related literature, yet not limited to specific tasks. In general, there are three main characteristics differentiating our work from existing techniques: being flexible to any deep networks, outlining a unified framework for modeling the bottom-up/top-down processes with end-to-end training, and being fully trainable and better interpretable with explicit inference processes. In general, our work is closely related to two streams of research in the literature:

Hierarchical graphical models have an enormous impact in computer vision, as they are powerful for expressing and capturing inherent structures, contextual information and high-level human knowledge. Their applications span from low-level problems, *e.g.*, hierarchical clustering, image restoration, to high-level tasks, *e.g.*, object parsing, human-object interaction. Commonly used models include MRF/CRF [13], part-based models [9, 18], and And-Or Graph [15, 28, 34]. For inference, bottom-up process passes information in a feed-forward manner while top-down process in a feed-back fashion over the hierarchy. In this work, we extend deep learning algorithm to hierarchical graphical model for end-to-end learning bottom-up and top-down processes jointly and automatically.

Deep learning with graphical models has recently received growing interests. Many recent works [5, 37, 38, 4] focus on incorporating CRF into networks with end-to-end training. Others extend RNN [22], or LSTM [27] from chain structures to tree or graph structures [29, 12, 16, 3]. However, they largely address the bottom-up process over structured architectures or work in a mixed fashion of bottom-up and top-down manners. In comparison, our model formulates the bottom-up and top-down inference processes in an explicit way. It is also a more principled and interpretable framework for modeling complex graph structures, rather than previous models limited to MRF assumptions or implicit mechanics.

Some works explore the **top-down mechanism in neural networks** and demonstrate success in their specific tasks. More specifically, bottom-up/top-down network architectures are proposed for leveraging both low-level and high-level features from different layers in semantic segmentation [21, 23]. Some investigations [2, 11] focus on

inspiring information flow between feed-forward and feedback loops in networks. However, these works (i) often perform inference over DNN hierarchy, without considering semantic hierarchical structures and relations in graph models; (ii) very few touch how to *explicitly* learn the bottom-up and top-down processes over a hierarchical graph. Additionally, the proposed model is more favored due to its post-hoc interpretability that specifies how its outputs are inferred from different information flows.

3. Representation

Given a graph $\mathcal{G}=(\mathcal{V},\mathcal{E})$, where nodes $\mathcal{V}$ represent problem components, and edges $\mathcal{E}$ capture the relationships between nodes (see Fig. 2(a)). Since we concentrate on hierarchical graphical model, we decompose nodes $\mathcal{V}$ into L layers: $\mathcal{V}=\mathcal{V}^1\cup\ldots\cup\mathcal{V}^L$, where $\mathcal{V}^l$ indicates the set of the nodes in l-th layer and the root node locates in the first layer (l=1). Edges can be further decomposed into three categories: $\mathcal{E}=\mathcal{E}^{com}\cup\mathcal{E}^{dec}\cup\mathcal{E}^{rel}$. $\mathcal{E}^{com}$ and $\mathcal{E}^{dec}$ are sets of vertical edges connecting parent nodes with their child nodes, which represent hierarchical constraints of composition and decomposition. Note vertical edges work in both bottom-up and top-down directions (*i.e.*, undirected edges), we use $\mathcal{E}^{com}$ and $\mathcal{E}^{dec}$ denote edging directing upwards and downwards, respectively. $\mathcal{E}^{rel}$ refers to the set of horizontal edges connecting among siblings with the same parent, which describes contextual relations in hierarchy. As suggested in [34], three inference processes, termed α, β and γ processes, can be derived for each node $v\in\mathcal{V}$.

α **process** detects node v directly based on image features. The α process is the basic inference, which can work alone (without taking advantage of surrounding context). Most structured networks [30, 5, 17, 4] in literature are proposed in this line. It can be viewed as either bottom-up or top-down. By bottom-up, it means that discriminative models. By top-down, it means that generative models are used.

β **process** computes node v by binding the detected child nodes in bottom-up fashion, where the child nodes' α processes are activated. An intuitive interpretation of this inference is to infer an occluded node, like human head, from its detected sub-nodes, say eye node or mouth node. Most component or part based models [16] belong to this process.

γ **process** predicts node v top-down from its parent nodes whose α processes are activated. The parent node passes contextual information, such as we can detect human head node even we only see the outline of the person. Most of the context-based methods belong to this process.

We propose a deep learning algorithm to learn above processes over the graph $\mathcal{G}$. In the high level, each node v is parameterized as a stack of CNNs by means of learning capacity and differentiable property. Each node accepts the information from other nodes, in bottom-up (β process) or top-down (γ process) manner as input, or directly uses the deep learning features from an underlying network (α process) for inference. In this way, we build a structured and fully differentiable network, which efficiently models $\mathcal{G}$ and explicitly learns inferences with powerful backpropagation.

4. Problem Formulation

According to Bayes rule, $\mathcal{G}$ can be solved by maximizing a posterior (MAP), that is,

$$\mathcal{G}^* = \arg\max_{\mathcal{G}} p(\mathcal{G}|I;\theta) \propto \arg\max_{\mathcal{G}} p(I|\mathcal{G};\theta)\cdot p(\mathcal{G};\theta), \tag{1}$$

where θ indicates the model parameters.

Likelihood $p(I|\mathcal{G};\theta)$ measures how well the observed image data I satisfies the hierarchical model $\mathcal{G}$. We assume each node in $\mathcal{G}$ only corresponds to a certain region of image, thus the likelihood $p(I|\mathcal{G};\theta)$ can be decomposed as:

$$p(I|\mathcal{G};\theta)=\prod_{v\in\mathcal{V}} p(I_{\Lambda_v}|v;\theta)=p^{bg}(I_\Lambda)\prod_{v\in\mathcal{V}}\frac{p^{fg}(I_{\Lambda_v}|v;\theta_v^\alpha)}{p^{bg}(I_{\Lambda_v})}, \tag{2}$$

where Λ denotes image lattice and Λ_v denotes the image region occupied by node v, $p^{bg}(\cdot)$ and $p^{fg}(\cdot)$ denote background and foreground probability, respectively. Similar to [26, 7], $p^{bg}(I_\Lambda)$ can be assumed as a constant and the likelihood ratio $g(\cdot)=p^{fg}(\cdot)/p^{bg}(\cdot)$ can be regarded as a logistic regression. This ratio represents the straight pass inference (*i.e.*, α **process**). For each node, the α process consists of two sub-steps: (i) extracting features ϕ_I from raw images and (ii) making direct predictions based on the extracted features.

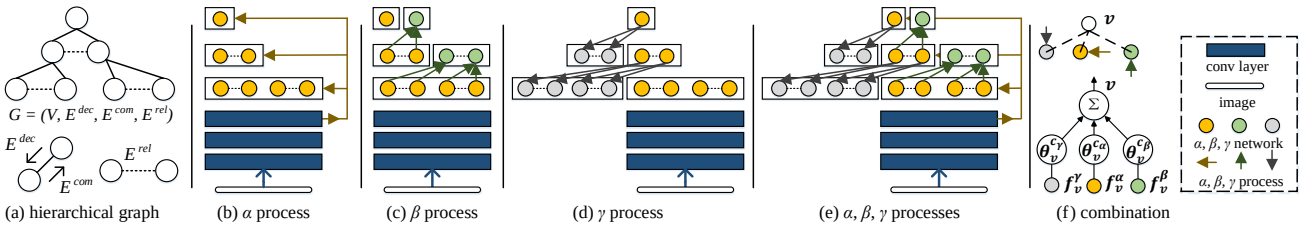
Prior $p(\mathcal{G};\theta)$ imposes constraints on the hierarchy, measuring the compatibilities among composition edges $\mathcal{E}^{com}$, decomposition edges $\mathcal{E}^{dec}$ and contextual edges $\mathcal{E}^{rel}$:

$$\begin{aligned} p(\mathcal{G};\theta)&=\prod_{v\in\mathcal{V}} p(v;\theta_v^c)\cdot p(nb(v)|v),\\ &=\prod_{v\in\mathcal{V}} p(v;\theta_v^c)\cdot p(ch(v)|v;\theta_v^\beta)\cdot p(pr(v)|v;\theta_v^\gamma)\cdot p(sb(v)|v), \end{aligned} \tag{3}$$

where $nb(v)$, $ch(v)$, $pr(v)$ and $sb(v)$ denote neighbors, children, parents, siblings of node v, respectively. Note that, for some nodes, the composition or decomposition edges (*i.e.*, β or γ processes) might not exist. Terminal leaf nodes only have α and γ processes, while the root node only has α and β processes. For clarity, we omit such cases, as they do not affect the method description.

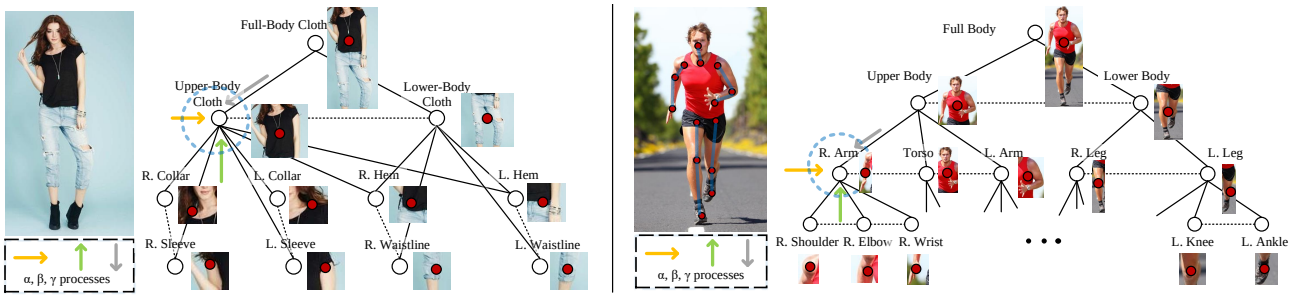
Prior term $p(v;\theta_v^c)$ measures to what extent we should trust different information sources (*i.e.*, information flows from α,β,γ processes). For each v, the fusion term $p(v;\theta_v^c)$ is defined as a weighted combination of α,β,γ processes:

$$\begin{aligned} &p(v;\theta_v^c)=[\theta_v^{c_\alpha},\theta_v^{c_\beta},\theta_v^{c_\gamma}],\\ &s.t.\ \theta_v^{c_\alpha}\geq 0,\ \theta_v^{c_\beta}\geq 0,\ \theta_v^{c_\gamma}\geq 0,\ \theta_v^{c_\alpha}+\theta_v^{c_\beta}+\theta_v^{c_\gamma}=1. \end{aligned} \tag{4}$$



(a) hierarchical graph (b) α process (c) β process (d) γ process (e) α, β, γ processes (f) combination

Figure 2. **The proposed α-β-γ network**. (a) The encoded hierarchical graph $\mathcal{G}$. (b)-(d) α-, β-, and γ- networks encoding α, β, γ process. (e) Joint inference based on neural network. (f) Fusion of three information flows. See text for detailed explanations.



(a) (b)

Figure 3. **Graphical representations for cloth landmark localization (a) and human pose estimation (b)**, where blue circles illustrate the α, β, γ processes of *upper-body cloth* node and *right arm* node, respectively.

Prior term $p(ch(v)|v;\theta_v^\beta)$ represents the bottom-up inference (*i.e.*, β **process**), which considers information flow upward from descendants. Each node v is fed with the information flow from its child nodes $ch(v)$, which composes composition edges $\mathcal{E}_v^{com}$.

Prior term $p(pr(v)|v;\theta_v^\gamma)$ represents the top-down inference (*i.e.*, γ **process**). Each node v is fed with the information flow from its parent nodes $pr(v)$ in γ process, which conveys the high-level information in a top-down manner. This describes decomposition edges $\mathcal{E}_v^{dec}$.

So far, we have discussed the formulation of nodes with vertical connections in the graph $\mathcal{G}$, which allows us to utilize information from straight pass (*i.e.*, α process), bottom-up process (*i.e.*, β process) and top-down process (*i.e.*, γ process). This generally covers composition relations $\mathcal{E}^{com}$ and decomposition relations $\mathcal{E}^{dec}$. Last but not least, our model should be able to capture contextual relations $\mathcal{E}^{rel}$.

Prior term $p(sb(v)|v)$ describes horizontal edges $\mathcal{E}^{rel}$ among siblings, which could represent many possible contextual relations, such as object-object interactions, dependency grammars and kinematic relations. In this paper, we consider contextual relations are encoded in prior terms $p(ch(v)|v;\theta_v^\beta)$ and $p(pr(v)|v;\theta_v^\gamma)$, which are joint distributions for child nodes and parent nodes given node v, respectively, while prior work usually assumes conditional independence among siblings. Thus we choose to implicitly model contextual relations in our α-β-γ network, which will be elaborated in next section.

In summary, our model encodes four probability distributions for each node v, parameterized by

$$\theta = \{(\theta_v^\alpha, \theta_v^\beta, \theta_v^\gamma, \theta_v^c) : v \in \mathcal{V}\}. \tag{5}$$

5. Learning

For each $v \in \mathcal{V}$, we further derive three sub-networks, namely α-, β-, and γ- network, for learning α, β, and γ processes.

α-process. The α-network f_v^α, parameterized by θ_v^α, is learnt for node v. It takes cropped images under corresponding lattice Λ_v as inputs and prediction score maps as outputs:

$$g(I_{\Lambda_v}|v;\theta_v^\alpha) = f_v^\alpha(\phi_{I_{\Lambda_v}};\theta_v^\alpha), \tag{6}$$

where the image features $\phi_{I_{\Lambda_v}}$ are extracted from a underlying network. As shown in Fig. 2(b), the final score map can be obtained by applying logistic *sigmoid* activation function.

β-process. As shown in Fig. 2(c), the β-network f_v^β for node v utilizes the information of its child nodes $ch(v)$ in α process, and outputs prediction score as the result of β process:

$$\begin{aligned} &p(ch(v)|v;\theta_v^\beta) \propto f_v^\beta(\phi_{ch(v)};\theta_v^\beta), \\ &\phi_{ch(v)} = \mathbf{P}_{avg}(\{f_{v'}^\alpha(v') : v' \in ch(v)\}), \end{aligned} \tag{7}$$

where we use channel-wise average-pooling operation $\mathbf{P}_{avg}$ for combining the output scores from child nodes. Such operation is important for transforming features from a variable number of predictions from child nodes to a fixed-size

feature representation. Note that any commutative operations can be used as alternatives (*e.g.*, sum-pooling, max-pooling).

γ-process. γ process works on the knowledge transferred from the parent node which is activated in α process. For node v, γ-network f_v^γ takes the information $\phi_{pr(v)}$ from parent nodes $pr(v)$ as input, and generates prediction as the output of γ process (see Fig. 2(d)):

$$\begin{aligned} &p(pr(v)|v;\theta_v^\gamma) \propto f_v^\gamma(\phi_{pr(v)};\theta_v^\gamma), \\ &\phi_{pr(v)} = \mathbf{P}_{avg}(\{f_{v'}^\alpha(v') : v' \in pr(v)\}). \end{aligned} \tag{8}$$

Fusion of α, β, γ processes. As illustrated in Fig. 2(f), the final prediction is made by a weighted combination of outputs generated from α-, β-, and γ- networks, parameterized by $\theta_v^c = [\theta_v^{c_\alpha}, \theta_v^{c_\beta}, \theta_v^{c_\gamma}]$. We represent the combination weights as 1×1 convolution layer connecting three channels (without bias term) and enforce the non-negativity and normalization constraints in Eqn.(4) to preserve model interpretability.

α-β-γ network. As shown in Fig. 2(e), the joint framework composes all the above components into a unified network, which can be learned in an end-to-end manner. Given ground-truth $\hat{v}$ for each node in the hierarchy $\hat{\mathcal{V}}$ with total K training samples, the α-β-γ network can be learned as:

$$\theta^* = \arg\max_\theta \prod_{k=1}^{K} p(\hat{\mathcal{V}}_k|I_k,\theta) = \arg\min_\theta \sum_{k=1}^{K} \sum_{\hat{v}\in\hat{\mathcal{V}}_k} L(\hat{v}|\theta_v), \tag{9}$$

where $L(\hat{v}|\theta_v)$ is the prediction loss for node v. The losses are defined as per the respective tasks, which are elaborated in next section. Our whole model is differentiable, and thus all the parameters θ of the graph model (in Eqn. 5) can be trained in a stochastic way.

Parameter sharing. Noticing that CNNs are naturally inherited to describe relations among all nodes in higher layer $\mathcal{V}^l$ and all nodes in lower layer $\mathcal{V}^{l+1}$. We thus employ parameter sharing among siblings to encourage information exchange, instead of modeling $\mathcal{E}^{rel}$ explicitly. We partition nodes $\mathcal{V}^l$ in l-th layer into N unconnected groups: $\mathcal{V}^l = \mathcal{V}_1^l \cup \ldots \cup \mathcal{V}_N^l$, according to their sibling relations. Then we enforce parameter sharing among nodes from same groups, instead of learning distinct parameters for each node. Taking Fig. 3(b) as an example, there exist kinematic relations (represented as dotted lines) among human body parts: *r. shoulder* $\leftrightarrow$ *r. elbow* and *r. elbow* $\leftrightarrow$ *r. wrist*, where three nodes are siblings with the same parent node *r. arm*. We model these three nodes with the same parameterization, which represents the knowledge sharing among them. Parameter sharing not only enables our network to capture complex inter-sibling relations and allows siblings to bootstrap each other capabilities, but also brings higher flexibility and better training efficiency on a large hierarchical structure. Overall, our entire model (including the underlying network) is fully differentiable, thus can be trained in end-to-end manner.

6. Experiments

We validate our α-β-γ network on two vision tasks: cloth landmark localization and human pose estimation. Then, we study post-hoc interpretability.

6.1. Cloth Landmark Localization

Fashion landmarks are functional keypoints defined on clothes, such as corners of neckline, cuff [20], which are effective representation for visual fashion understanding. Cloth landmark detection is a good example with inherent structures and obvious components, yet challenging due to background clutters, deformations, and scales.

Dataset. We use Fashion Landmark Detection (FLD) [20][1], which contains totally $123,016$ clothes images. For each image, 8 fashion landmarks (*l.&r. collar*, *l.&r. sleeve*, *l.&r. waistline*, *l.&r. hem*) are annotated. For each image, cloth bounding box is also annotated.

Network architecture. A three-layer graph $\mathcal{G}$ is derived for representing human cloth (Fig. 3 (a)). We build our structured network following $\mathcal{G}$. The first five convolutional stacks of ResNet50 [10] are opted as our underlying network. For preserving detailed spatial information, we modify the last two blocks by changing the strides to 1. Specifications of α-, β-, and γ- networks are listed in Table 1. Note that 1×1 convolution layer with *sigmoid* layer is applied to produce final predictions. The principle behind such design is mainly for pursuing large enough receptive field and simplicity. The input images are resized into 224×224. Thus if sliding our network over the input image, we could obtain a 28×28 prediction map of each nodes for their specific tasks.

Training. For all nodes, ground-truth heatmaps are generated by convolving binary annotation maps with a small Gaussian kernel. For those higher-layer nodes without annotation, such as *upper-body cloth* or *full-body cloth*, we generate their annotation according to child nodes' configurations. This annotation process is similar to [18]. For node v, we would have an output prediction score map $S \in [0,1]^{28\times28}$ and its corresponding ground-truth map $\hat{S} \in [0,1]^{28\times28}$ for a 224×224 training image. Then we adopt Kullback-Leibler Divergence to measure the loss:

$$L(\hat{v}|\theta_v) = D_{KL}(S_v, \hat{S}_v) = \mathbf{1}(\hat{v}) \cdot \sum\nolimits_k^{28\times28} \hat{s}_k \cdot \log \frac{\hat{s}_k}{s_k}. \tag{10}$$

where the indicator function $\mathbf{1}(\cdot)$ is employed for remedying missing ground truth locations of the landmarks, in the sense that the error is not propagated back when a landmark

[1] Available at http://mmlab.ie.cuhk.edu.hk/projects/DeepFashion/LandmarkDetection.html

Nodes	Configuration
α-network	conv(256,(3,3)) $\rightarrow$ conv(128,(5,5)) $\rightarrow$ conv(64,(3,3)) $\rightarrow$ conv(32,(3,3))
β-network	conv(32, (7,7)) $\rightarrow$ conv(32, (7,7)) $\rightarrow$ conv(32,(7,7)) $\rightarrow$ conv(32,(7,7)) $\rightarrow$ conv(32,(5,5))
γ-network	conv(32, (7,7)) $\rightarrow$ conv(32, (7,7)) $\rightarrow$ conv(32,(7,7)) $\rightarrow$ conv(32,(7,7)) $\rightarrow$ conv(32,(5,5))

Table 1. **Configurations of α-, β-, and γ- networks**. Keras notations (channels, kernel) are used to define the conv layers.

Methods	3rd Layer								2nd Layer		1st Layer
	L.Collar	R.Collar	L.Sleeve	R.Sleeve	L.Waistline	R.Waistline	L.Hem	R.Hem	U.Body	L.Body	F.Body
FashionNet CVPR'16 [19]	.0784	.0803	.0975	.0923	.0874	.0821	.0802	.0893	-	-	-
DFA ECCV'16 [20]	.048	.048	.091	.089	-	-	.071	.072	-	-	-
DLAN AAAI'17 [35]	.0531	.0547	.0705	.0735	.0752	.0748	.0693	.0675	-	-	-
FashionGrammar CVPR'18 [32]	.0463	.0471	.0627	.0614	.0635	.0692	.0635	.0527	-	-	-
α-network	.0457	.0450	.0619	.0628	.0623	.0705	.0643	.0530	.1220	.1186	.0935
β-network	-	-	-	-	-	-	-	-	.1100	.1105	.0820
γ-network	.0503	.0512	.0721	.0713	.0643	.0821	.0703	.0627	.1002	.1013	-
α-β-γ network w/o share	.0441	**.0415**	.0606	.0615	.0620	.0702	**.0624**	.0515	.0994	.0986	.0790
α-β-γ network	**.0435**	.0426	**.0597**	**.0612**	**.0614**	**.0690**	.0631	**.0511**	**.0989**	**.0977**	**.0778**

Table 2. **Comparison of normalized error (NE) on FLD dataset**. Lower values are better. The best score is marked in **bold**.

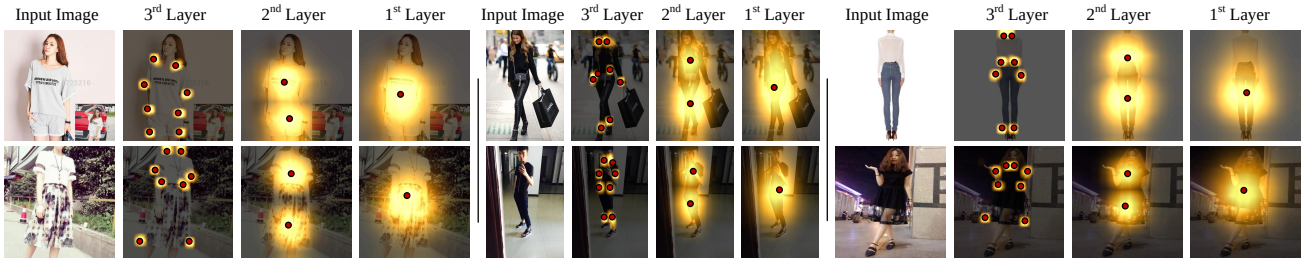


Figure 4. **Results of cloth landmark localization**. We show the prediction scores of each layer in our hierarchical graph, where the brighter pixel indicates higher prediction values, and the red circle indicates the location of highest score of each node.

is occluded (according to the visibility annotation). Here we drop subscript v for s_k and $\hat{s}_k$ for simplicity.

Performance comparison. We compare α-β-γ network with four deep learning based fashion landmark detectors: [19, 20, 35, 32]. For all the methods, standard train/validation/test settings $(83,033/19,992/19,991)$ in FLD dataset are used for fair comparisons. We adopt normalized error (NE) metric suggested by FLD dataset for evaluation. NE refers to the ℓ_2 distance between predicted landmarks and ground-truth in the normalized coordinate space (*i.e.*, divided by the width/height of the image). We report the results in Table 2, where the baselines: α-*network*, β-*network*, and γ-*network* indicate the results obtained from α, β, and γ processes independently. α-β-γ *network w/o share* corresponds to the results of α-β-γ network without parameter sharing, equivalent to ignoring the horizontal relations $\mathcal{E}_{rel}$ in graph $\mathcal{G}$. As seen, the proposed structured network outperforms other competitors. Some qualitative results can be found in Fig. 4.

Discussion. The improvement would be attributed to the integration of deep learning and graph model. Unstructured models like FashionNet and DFA are hard to model the inherent structures of fashion cloth, which offers strong contextual information about cloth landmark locations. Our solution is more favored due to its structural modeling with underlying graphical representation and powerful joint bottom-up and top-down inference. We can further observe that, α inference performs better for those low-level nodes (*e.g.*, *l. collar*, *r. waistline*), while β and γ processes are more informative for high-level nodes like *lower-body cloth* or *full-body cloth*. Compared with those explicit junctions, the nodes in higher layers are often companied with more ambiguities, in which sense more complex bottom-up/top-down inference processes are preferred. For α-β-γ *network w/o share*, we can observe a drop of performance. This demonstrates the importance of structure information, and thus verifies our design. Besides, parameter sharing would bring extra advantage of better generalization.

6.2. Human Pose Estimation

In this section, we present our structured network for another vision task, human pose estimation, which is a popular vision task requiring both powerful detection of human body parts and effective modeling of relationship among parts. We compare our methods with several previous pose estimators using graphical structures or pure networks, which show the benefits of capturing the interactions between body parts with graphical models. In this experimen-

Methods	4th Layer							3rd Layer			2nd Layer		1st Layer
	Head &Neck	Shoulder (L.&R.)	Elbow (L.&R.)	Wrist (L.&R.)	Hip (L.&R.)	Knee (L.&R.)	Ankle (L.&R.)	Arm (L.&R.)	Leg (L.&R.)	Head &Torso	U.Body	L.Body	F.Body
Wang et al. CVPR'13 [31]	89.1	78.5	62.5	52.3	85.2	69.6	65.9	-	-	-	-	-	-
Chen et al. NIPS'14 [6]	91.8	78.2	71.8	65.5	73.3	70.2	63.4	-	-	-	-	-	-
Tompson et al. NIPS'14 [30]	90.6	79.2	67.9	63.4	69.5	71.0	64.2	-	-	-	-	-	-
Fan et al. CVPR'15 [8]	92.4	75.2	65.3	64.0	75.7	68.3	70.4	-	-	-	-	-	-
Yang et al. CVPR'16 [36]	90.6	78.1	73.8	68.8	74.8	69.9	58.9	-	-	-	-	-	-
Hu et al. CVPR'16 [11]	93.4	83.2	77.3	72.1	87.6	79.6	76.8	-	-	-	-	-	-
Wei et al. CVPR'16[33]	94.1	86.0	78.9	76.0	88.7	82.3	77.4	-	-	-	-	-	-
Liu et al. AAAI'16 [18]	88.4	76.5	70.6	66.3	75.6	68.7	67.5	67.4	63.2	73.4	74.7	76.2	78.9
Park et al. PAMI'18 [24]	90.7	79.8	76.8	68.1	78.9	70.2	75.4	76.9	75.1	82.5	84.9	82.0	81.5
α-network	94.4	**86.4**	79.3	75.2	89.1	82.6	77.1	78.4	76.2	83.4	89.1	85.5	84.7
β-network	-	-	-	-	-	-	-	77.6	78.3	83.5	90.7	92.5	91.7
γ-network	92.5	82.1	76.5	71.1	85.7	78.3	72.2	80.3	81.0	82.4	90.2	90.3	-
α-β-γ-network w/o share	95.1	85.0	80.6	**78.5**	90.3	83.2	78.6	80.7	79.3	84.2	91.3	93.4	92.4
α-β-γ network	**95.6**	85.3	**81.6**	77.3	**91.3**	**83.7**	**80.5**	**82.5**	**81.3**	**86.5**	**93.4**	**94.9**	**92.8**

Table 3. **Comparison of PCKh metric on LSP dataset**. Higher values are better. The best score is marked in **bold**.

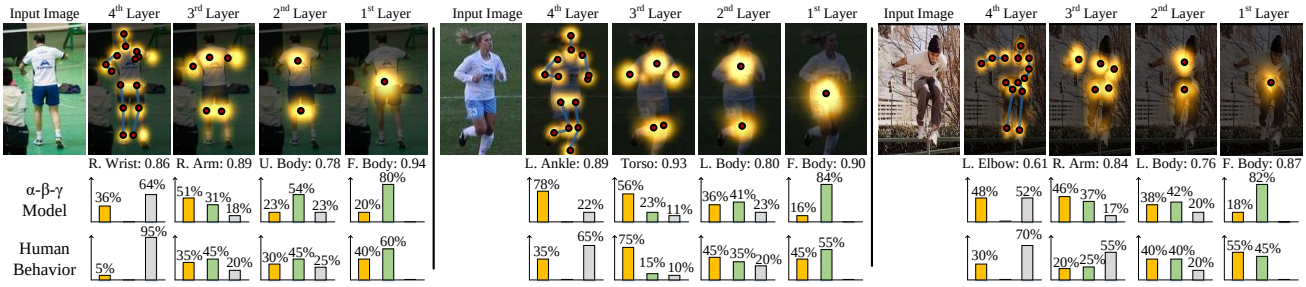


Figure 5. **Results of human pose estimation on LSP dataset**. In the first row, we show the predictions of each layer. For each layer, we select one node to demonstrate its prediction score. Then, in the second and third rows, we present the contributions of α, β, and γ processes over such node, which estimated from our model and human behavior.

t, we demonstrate our structured network is able to generate more accurate pose estimations via explicitly and jointly considering bottom-up and top-down information.

Dataset. We use the standard pose estimation benchmark: LSP dataset [14][2], containing $11,000$ images for training and $1,000$ images for testing. The images are of people in various sport poses.

Network architecture. In Fig. 3 (b), human pose is represented as a 4-level hierarchical structured network. The bottom level of our hierarchy is comprised of the 14 atomic parts corresponding to the annotated joints. The third level consists of 5 composite parts formed by grouping parts belonging to each of the limbs, a composite part for the head and torso. The second-layer nodes refer to the upper-human body and lower-human body, and the root node presents full-human body. Such settings are similar to previous graphical models [18, 25]. The base network of our model is built upon [33]. For consistency, we adopt the same network architectures of α-, β-, and γ- networks as in Table 1.

Training. We follow the standard protocol in the area of pose estimation. For each node, a ground-truth confidence map is created by putting Gaussian peaks at ground-truth locations of corresponding part. We infer the ground-truth locations of higher-level parts following the annotation procedure in fashion landmark detection. We also resize the input images into 224×224. Then we use ℓ_2 distance to measure the loss, which is widely used in human pose estimation:

$$L(\hat{v}|\theta_v) = D_{\ell_2}(S_v, \hat{S}_v) = \sum_k^{28\times 28} \|s_k - \hat{s}_k\|_2, \quad (11)$$

where the $S \in [0,1]^{28\times 28}$ and $\hat{S} \in [0,1]^{28\times 28}$ denote the output score map and the ground-truth, respectively.

Performance comparison. For evaluation, we use the PCKh metric [1], which is a modification of the Percentage Correct Keypoints (PCK) metric with a matching threshold. We compare the performance of our method with several pose estimators. We also investigate the performance of individual inference processes, and simplified model without horizontal relations. As seen in Table 3, our model outperforms other competitors. We visualize detection results in Fig. 5.

Discussion. The proposed model offers a powerful tool that has the complementary strengths of neural network and

[2]Available at www.comp.leeds.ac.uk/mat4saj/lspet.html

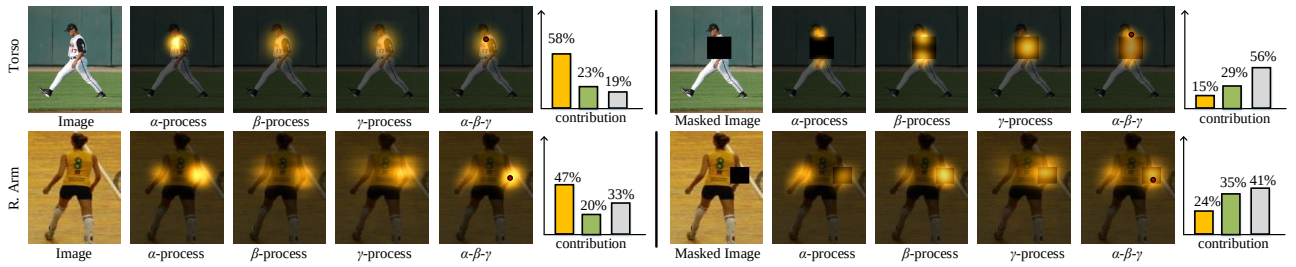


Figure 6. **Examining intrepretability with masked examples**. See text for more details.

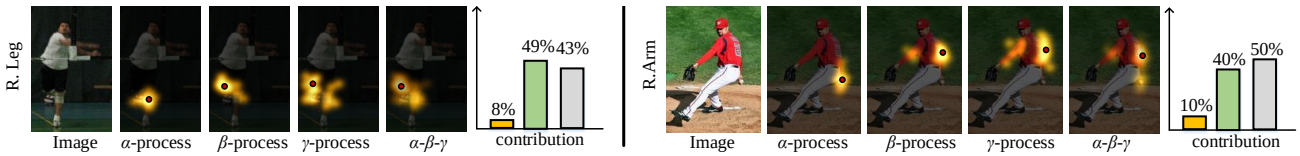


Figure 7. **Illustration of bottom-up and top-down inference**. We select one node in the 3*rd* layer of human pose graph and show the predictions from α-, β-, and γ- processes, and draw the distribution of contribution of above processes in the final prediction.

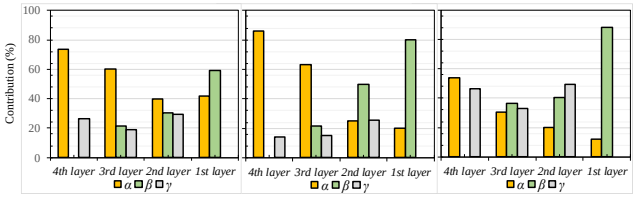
graphics models. It is not limited to CRF-like assumptions, which are widely used in previous graphical pose models. Therefore, α-β-γ network is able to better represent rich internal relations among human body parts. When comparing the performance of individual inference process and our full model, we again get the similar observations that α, β, and γ processes are favored under different scenarios and the integration of three processes would improve final performance.

6.3. Study of Post-hoc Interpretability

We further explore the post-hoc interpretability conferred by our model, specifically, how the contribution made by different inference processes coincides with human knowledge. The contribution of a process is defined as the ratio between its own weighted prediction and the final score. Taking α process as an example, the contribution of α process for node v can be formulated as:

$$C^{\alpha}(v) = \frac{\theta_v^{c_\alpha} \cdot f_v^{\alpha}(v)}{\theta_v^{c_\alpha} \cdot f_v^{\alpha}(v) + \theta_v^{c_\beta} \cdot f_v^{\beta}(v) + \theta_v^{c_\gamma} \cdot f_v^{\gamma}(v)}. \quad (12)$$

We first perform a user study to measure the agreement between human behavior and our model. A corpus of 20 participants (9 female) with diverse backgrounds are recruited to participate in our studies. 100 images were randomly selected from the test set of LSP dataset. For each node, participants were asked to label the most informative inference process. Generally, for the cases that a node can be directly recognized, it is labeled as α label. In the situation that compositional or contextual information are needed, β or γ are annotated accordingly. We average the votes from all the participants as human consensus. Fig. 8 (a-b) plot the contribution distributions of the three processes annotated from human and learned via our model, showing



(a) (b) (c)

Figure 8. **Numerical study of contributions of three inference processes**: (a) human behavior; (b) performance of α-β-γ network; and (c) α-β-γ network with masked images. We average the scores from same-layer nodes.

that α process is important in low-level nodes, while β and γ processes are relatively strong in high-level nodes. This observation is also verified in previous experiments, that α, β and γ processes are effective in different layers. We also find the contribution distribution of our model is close to human consensus.

In Fig. 7, we select one node in the 3*rd* layer of human pose graph and show the prediction scores from α, β, and γ processes and the final score from the fusion of above three processes. We also present the distribution of contribution of above processes in the final prediction. As seen, the combination of α, β, and γ inference processes would get the best results. Quantitatively, our model obtains 61.3%, 50.7% and 31.5% *average precision* (AP) of α, β, and γ processes with human consensus over all nodes, respectively.

We further conduct a counterfactual experiment using data manipulation. For each image, we generate a mask (20×20) to cover certain nodes. Afterwards, we obtain manipulated images with occlusions on body parts and re-estimate human poses (see Fig. 6). The statistics regarding contributions of three processes are reported in Fig. 8 (c).

Interestingly, we find that, when a node suffers occlusion, β and γ processes provide more supports in final predictions.

7. Conclusion

We propose a deep structured network for combining hierarchical graph representations with deep learning. The α-β-γ network is capable of modeling rich structures, incorporating top-down/bottom-up inference learned in end-to-end manner. It gains better interpretability via separating explicit inferences from the underlying implicit mechanics of neural network. Performance and interpretability of the proposed model are well demonstrated through extensive experiments on fashion landmark detection and human pose estimation.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *ECCV*, 2014. 7

[2] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu, D. Ramanan, and T. S. Huang. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *ICCV*, 2015. 2

[3] S. Cao, W. Lu, and Q. Xu. Deep neural networks for learning graph representations. In *AAAI*, 2016. 2

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *arXiv preprint arXiv:1606.00915*, 2016. 2, 3

[5] L.-C. Chen, A. Schwing, A. Yuille, and R. Urtasun. Learning deep structured models. In *ICML*, 2015. 2, 3

[6] X. Chen and A. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *NIPS*, 2014. 7

[7] Y. Chen, L. Zhu, A. Yuille, and H. Zhang. Unsupervised learning of probabilistic object models (poms) for object classification, segmentation, and recognition using knowledge propagation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(10):1747–1761, 2009. 3

[8] X. Fan, K. Zheng, Y. Lin, and S. Wang. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In *CVPR*, 2015. 7

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, 2010. 2

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[11] P. Hu and D. Ramanan. Bottom-up and top-down reasoning with hierarchical rectified gaussians. In *CVPR*, 2016. 2, 7

[12] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena. Structural-RNN: Deep learning on spatio-temporal graphs. In *CVPR*, 2016. 2

[13] T. Joachims, T. Finley, and C.-N. J. Yu. Cutting-plane training of structural SVMs. *Machine Learning*, 77(1):27–59, 2009. 2

[14] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 7

[15] I. Kokkinos, P. Maragos, and A. Yuille. Bottom-up & top-down object detection using primal sketch features and graphical models. In *CVPR*, 2006. 2

[16] X. Liang, L. Lin, X. Shen, J. Feng, S. Yan, and E. P. Xing. Interpretable structure-evolving LSTM. In *CVPR*, 2017. 2, 3

[17] G. Lin, C. Shen, A. van den Hengel, and I. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *CVPR*, 2016. 3

[18] J. Liu, Y. Li, P. Allen, and P. N. Belhumeur. Articulated pose estimation using hierarchical exemplar-based models. In *AAAI*, 2016. 2, 5, 7

[19] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 6

[20] Z. Liu, S. Yan, P. Luo, X. Wang, and X. Tang. Fashion landmark detection in the wild. In *ECCV*, 2016. 5, 6

[21] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2

[22] L. Medsker and L. C. Jain. *Recurrent neural networks: design and applications*. CRC press, 1999. 2

[23] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *ICCV*, 2015. 2

[24] S. Park, X. Nie, and S.-C. Zhu. Attribute and-or grammar for joint parsing of human pose, parts and attributes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018. 7

[25] V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell, and Y. Sheikh. Pose machines: Articulated pose estimation via inference machines. In *ECCV*, 2014. 7

[26] H. Schneiderman and T. Kanade. Object detection using the statistics of parts. *Int. J. Comput. Vis.*, 56(3):151–177, 2004. 3

[27] R. Socher, C. C. Lin, C. Manning, and A. Y. Ng. Parsing natural scenes and natural language with recursive neural networks. In *ICML*, 2011. 2

[28] X. Song, T. Wu, Y. Jia, and S.-C. Zhu. Discriminatively trained and-or tree models for object detection. In *CVPR*, 2013. 2

[29] K. S. Tai, R. Socher, and C. D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *ACL*, 2015. 2

[30] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, 2014. 3, 7

[31] F. Wang and Y. Li. Beyond physical connections: Tree models in human pose estimation. In *CVPR*, 2013. 7

[32] W. Wang, Y. Xu, J. Shen, and S.-C. Zhu. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *CVPR*, 2018. 6

[33] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016. 7

[34] T. Wu and S.-C. Zhu. A numerical study of the bottom-up and top-down inference processes in and-or graphs. *Int. J. Comput. Vis.*, 93(2):226–252, 2011. 1, 2, 3

[35] S. Yan, Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Unconstrained fashion landmark detection via hierarchical recurrent transformer networks. In *AAAI*, 2017. 6

[36] W. Yang, W. Ouyang, H. Li, and X. Wang. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *CVPR*, 2016. 7

[37] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015. 2

[38] X. Zhu, P. Sobihani, and H. Guo. Long short-term memory over recursive structures. In *ICML*, 2015. 2